Sentiment Analysis of Four Medical Conditions

George Fisher

Thursday, July 17, 2014

Summary

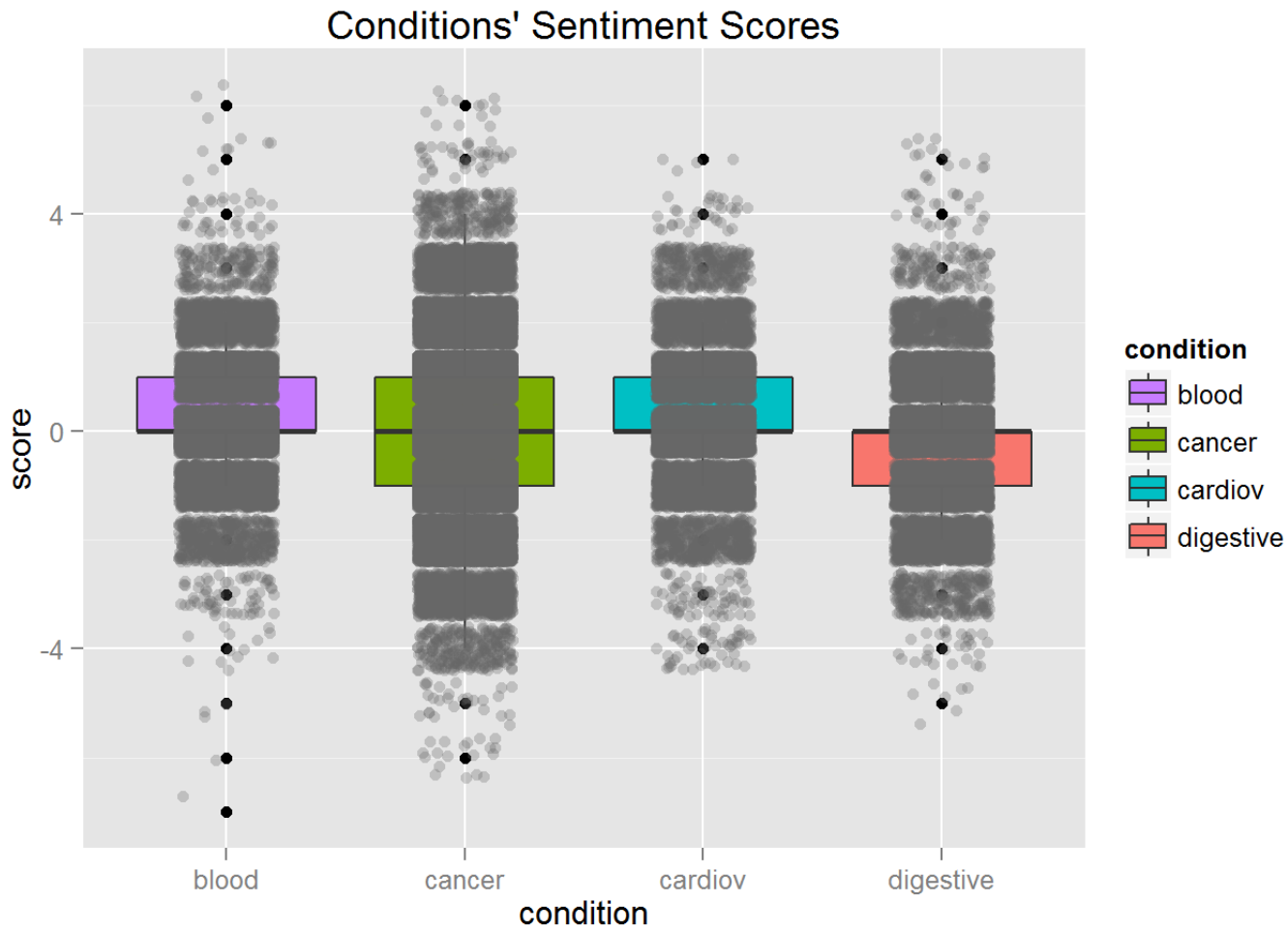
Using the Jeff Breen sentiment scoring system, we look at the sentiment expressed by people tweeting about four medical conditions: Blood Disorders, Cancers of all sorts, Cardiovascular Conditions and Digestive Disorders.

It is clear from this analysis that among these four, Digestive Disorders have the most negative effect on those who tweet about these diseases.

Distribution of sentiment: Boxplot

The dark gray dots represent the individual data points, 14,000 per condition.

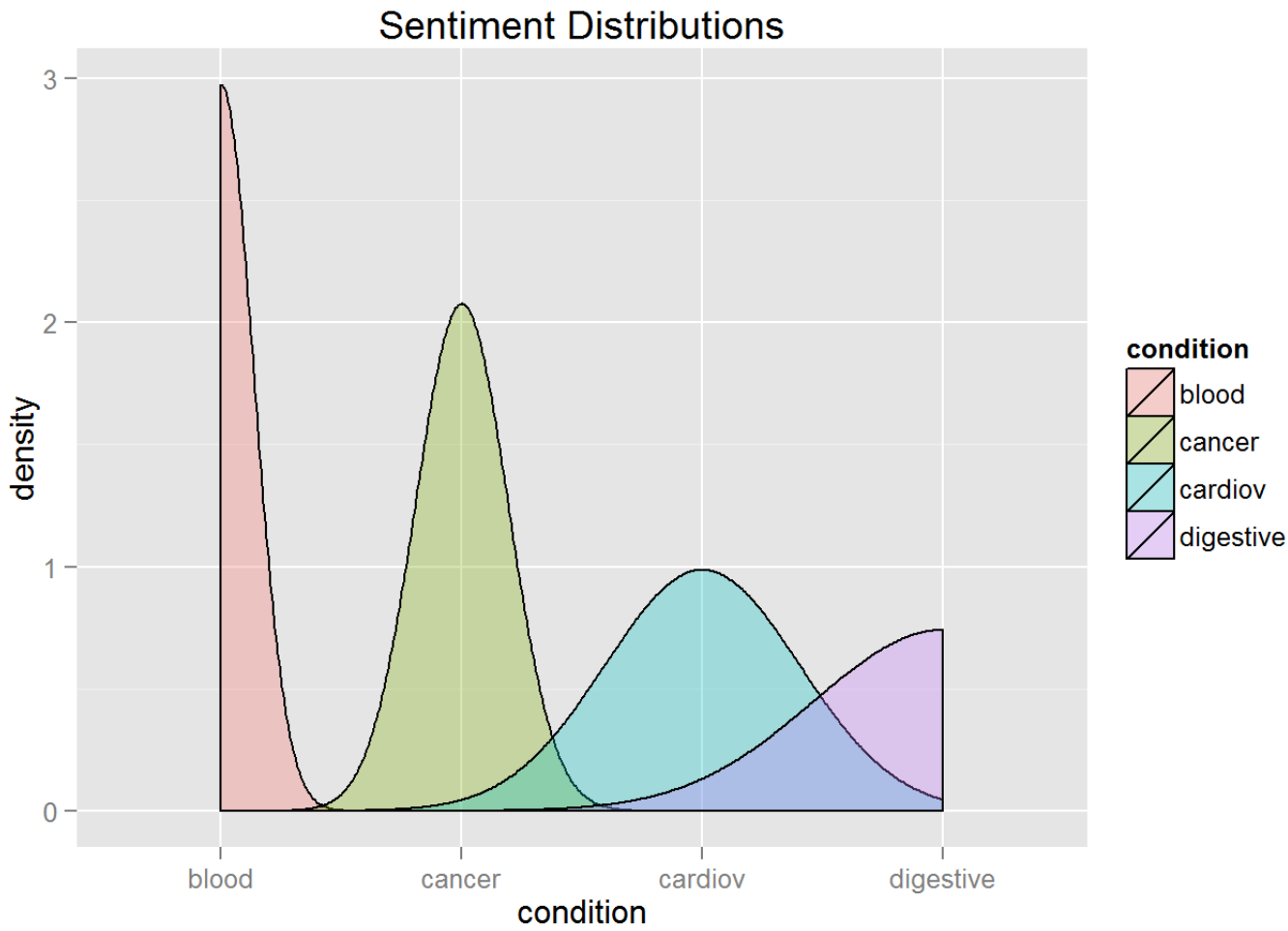
The boxplots in color represent the distribution of the sentiment for each condition. They all have their median nearly at zero with a very wide dispersion in both the positive and negative direction. Blood and Cardiovascular disorders seem to be somewhat skewed toward positve overall sentiment, while Digestive disorders are skewed toward the negative ranges.



Distribution of sentiment: Histogram

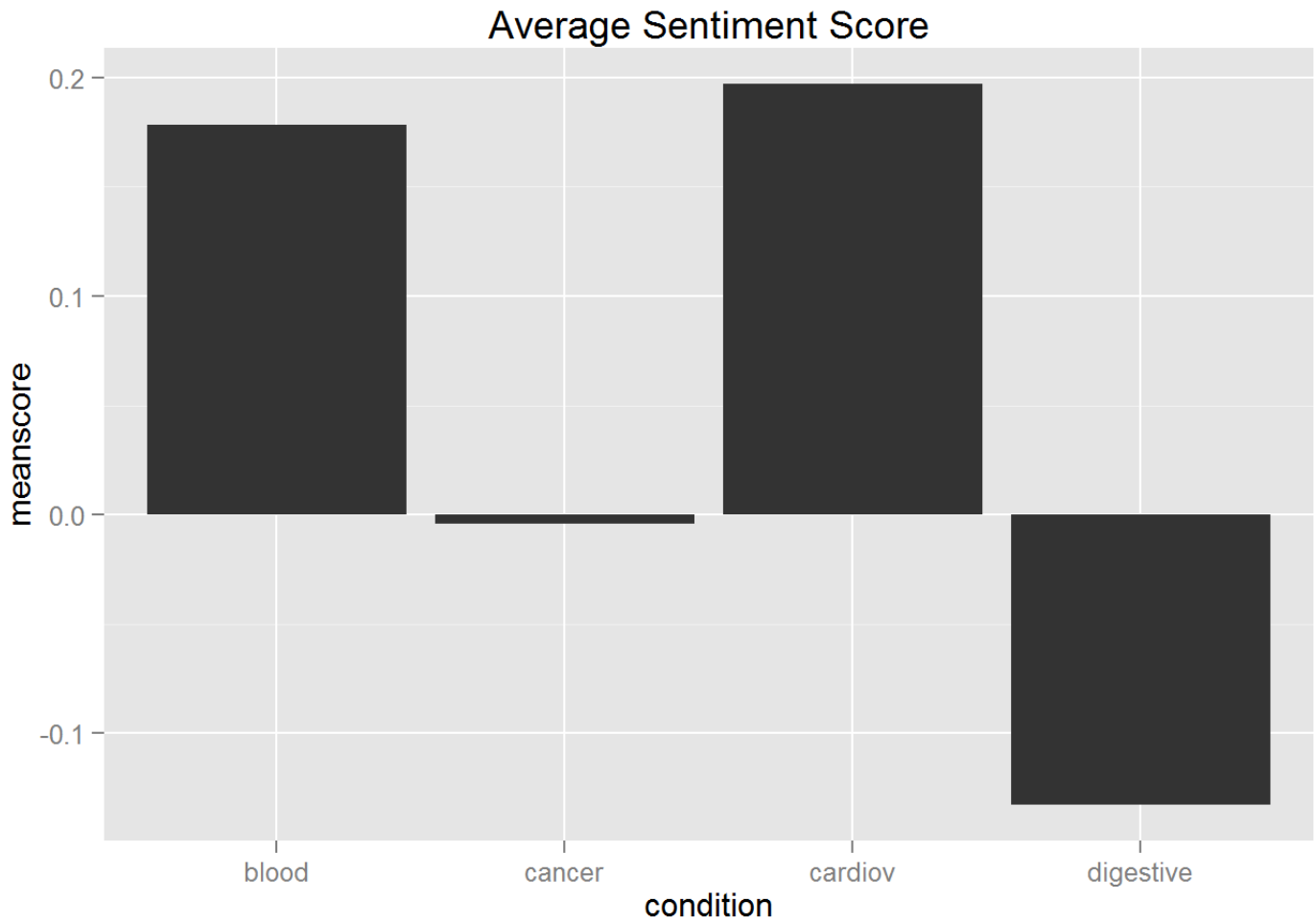
Another way to look at the distribution of sentiment is to show a smoothed histogram. For each condition, the vertical white line over the label is the average for that category and the plot shows the distribution although the left-tail of Blood and the right tail of Digestive are not plotted.

Blood is in a tight range around its mean, while Digestive has the greatest spread.



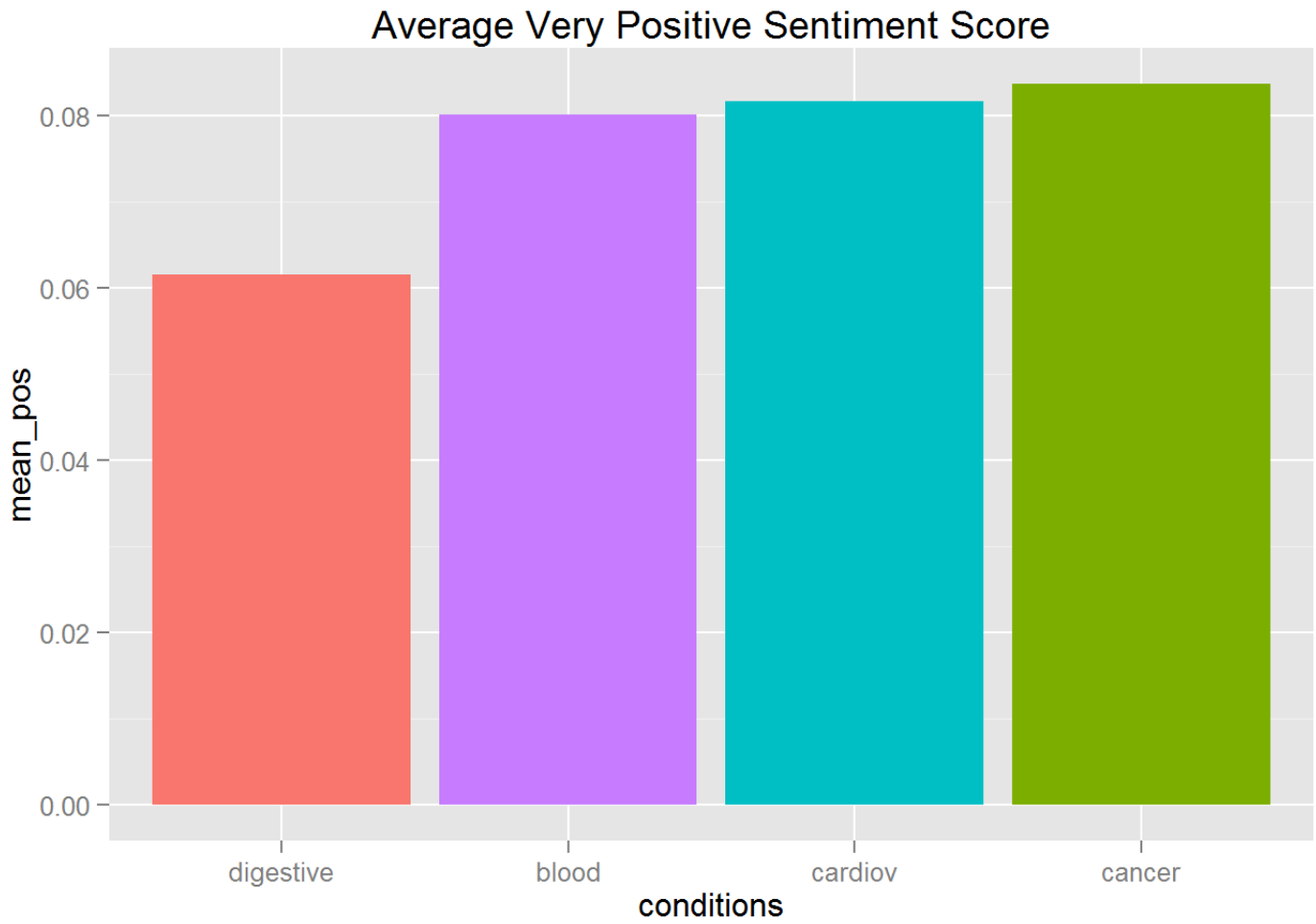
Average Scores

The averages show us very starkly what we saw in the distributions: Digestive disorders have by far the most negative effect on their suffers.



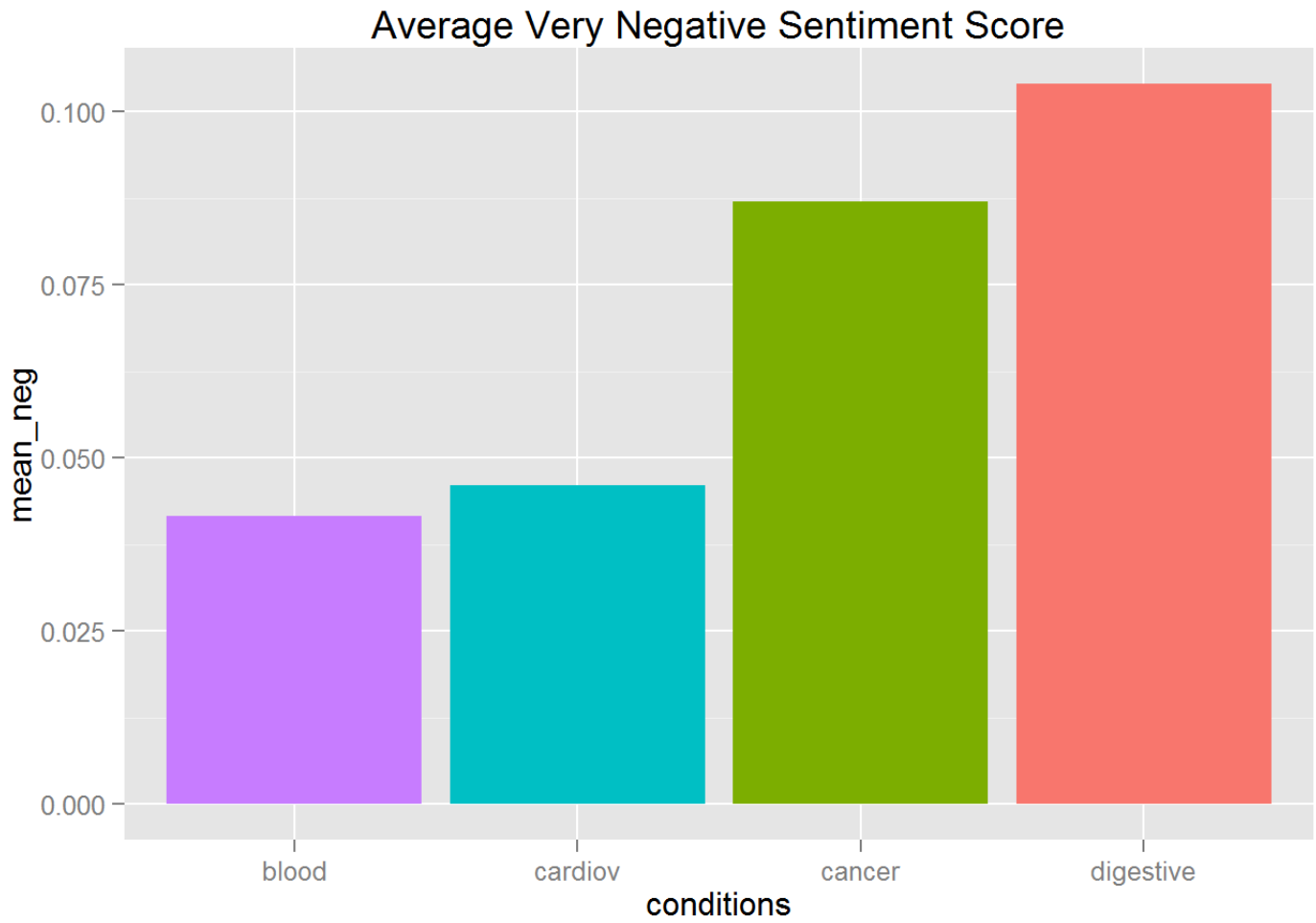
Positive Scores

More reinforcement for what we have already seen: Digestive disorders have a negative psychological effect to the extent of having the lowest positive scores.



Negative Scores

The same story again: none are good, but of these four tweets about Digestive Disorders show the greatest tendency toward negativity.



Credits and Description

```
# ============================================================================
# Title:         Sentiment_Analysis_Sanchez.R
# Author:        Gaston Sanchez
# Date:          May, 2012
#
# Modified By:   George Fisher
# Modified Date:July 2014
#
# Description:   Script showing how to perform a sentiment analysis based on
#                the approach described by Jeffrey Breen in his
#                "twitter text mining R slides"
#
# Modification: Instead of pulling live tweets of drinks we
#                read the csv files I created for the Healthcare
#                Twitter Analysis project.
#                What Sanchez did was take Jeffry Breen's work
#                and adapt it. I, in turn, have modified Sanchez's work.
#
# License:       BSD Simplified License
```

```
#               http://www.opensource.org/license/BSD-3-Clause
#               Copyright (c) 2012, Gaston Sanchez
#               All rights reserved
# ========================================================================



# ------------------------------------------------------------------------
# The general idea is to calculate a sentiment score for each tweet so we
# can know how positive or negative is the posted message.
# There are different ways to calculate such scores, and you can even
# create your own formula.
# We'll use a very simple yet useful approach to define our score formula
#   Score  =  Number of positive words  -  Number of negative words
# If Score > 0, then the sentence has an overall 'positive opinion'
# If Score < 0, then the sentence has an overall 'negative opinion'
# If Score = 0, then the  sentence is considered to be a 'neutral opinion'
#
# In order to count the number of positive and negative words,
# we need a very important ingredient:
# an opinion lexicon in english, which fortunately it is provided
# by Hu and Liu and it can be accessed from:
# http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
#
# Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews."
# Proceedings of the ACM SIGKDD International Conference on Knowledge
# Discovery and Data Mining (KDD-2004),
# Aug 22-25, 2004, Seattle, Washington, USA,

# Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing
# and Comparing Opinions on the Web." Proceedings of the 14th
# International World Wide Web conference (WWW-2005)
# May 10-14, 2005, Chiba, Japan.
# ------------------------------------------------------------------------
```

Programming Environment Details

```
sessionInfo()
```

```
## R version 3.1.0 (2014-04-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_1.0.0 stringr_0.6.2 plyr_1.8.1
##
## loaded via a namespace (and not attached):
##  [1] colorspace_1.2-4 digest_0.6.4     evaluate_0.5.5   formatR_0.10
##  [5] grid_3.1.0       gtable_0.1.2     htmltools_0.2.4  knitr_1.6
##  [9] labeling_0.2     MASS_7.3-33      munsell_0.4.2    proto_0.3-10
## [13] Rcpp_0.11.2      reshape2_1.4     rmarkdown_0.2.46 scales_0.2.4
## [17] tools_3.1.0      yaml_2.1.13
```